CSE 515T (Spring 2015) Assignment 1 Solutions

1. (Barber.) Suppose that a study shows that 90% of people who have contracted Creutzfeldt–Jakob disease (“mad cow disease”) ate hamburgers prior to contracting the disease. Creutzfeldt–Jakob disease is incredibly rare; suppose only one in a million people have the disease.

If you eat hamburgers, should you be worried? Does this depend on how many other people eat hamburgers?

Solution

Let $CJ$ be the random variable “has Creutzfeldt–Jakob disease,” and let $H$ be the random variable “eats hamburgers.” From the problem, we know $\Pr(H \mid CJ) = 0.9$ and $\Pr(CJ) = 10^{-6}$. From Bayes’ theorem, we may compute the posterior probability of having Creutzfeldt–Jakob disease given that you eat hamburgers:

$$\Pr(CJ \mid H) = \frac{\Pr(H \mid CJ) \Pr(CJ)}{\Pr(H)}.$$

The denominator, $\Pr(H)$, is the probability that a random person eats hamburgers, so whether you should be worried does depend on this value. I estimate $\Pr(H)$ might be around $\frac{1}{2}$; plugging in this value, the posterior probability of having Creutzfeldt–Jakob disease is

$$\frac{0.9 \times 10^{-6}}{1/2} = 1.8 \times 10^{-6},$$

so your probability of having the disease has only increased from $10^{-6}$ to $1.8 \times 10^{-6} = 0.0000018$ as a result of eating hamburgers. You can sleep safe.

If hamburger eating were rare, say $\Pr(H) = 10^{-5}$, then we would instead have $\Pr(CJ \mid H) = 9\%$, and maybe you should be worried!
2. (O’Hagan and Forster.) Suppose \( x \) has a Poisson distribution with unknown mean \( \theta \):

\[
p(x \mid \theta) = \frac{\theta^x}{x!} \exp(-\theta), \quad x = 0, 1, \ldots
\]

Let the prior for \( \theta \) be a gamma distribution:

\[
p(\theta \mid \alpha, \beta) = \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta \theta), \quad \theta > 0
\]

where \( \Gamma \) is the gamma function. Show that, given an observation \( x \), the posterior \( p(\theta \mid x, \alpha, \beta) \) is a gamma distribution with updated parameters \( (\alpha', \beta') = (\alpha + x, \beta + 1) \).

**Solution**

From Bayes’ theorem, we have:

\[
p(\theta \mid x) \propto p(x \mid \theta)p(\theta)
\]

\[
\propto (\theta^x \exp(-\theta))(\theta^{\alpha-1} \exp(-\beta \theta))
\]

\[
= \theta^{x+\alpha-1} \exp(-(\beta + 1)\theta)
\]

\[
\propto \mathcal{G}(\alpha + x, \beta + 1).
\]

Here we exploit a common trick: we manipulate the numerator, ignoring constants independent of \( \theta \). If we can recognize the functional form as belonging to a distribution family we know, we can simply identify the parameters and trust that the distribution normalizes!
3. (Optimal Price is Right bidding.) Suppose you have a standard normal belief about an unknown parameter $\theta$, $p(\theta) = \mathcal{N}(\theta; 0, 1^2)$. You are asked to give a point estimate $\hat{\theta}$ of $\theta$, but are told that there is a heavy penalty for guessing too high. The loss function is

$$
\ell(\hat{\theta}, \theta; c) = \begin{cases} 
(\theta - \hat{\theta})^2 & \hat{\theta} < \theta; \\
c & \hat{\theta} \geq \theta,
\end{cases}
$$

where $c > 0$ is a constant cost for overestimating. What is the Bayesian estimator in this case? How does it change as a function of $c$?

**Solution**

Let $\phi(x) = \mathcal{N}(x; 0, 1^2)$ be the standard normal pdf evaluated at $x$, and let $\Phi(x) = \int_{-\infty}^{x} \phi(x) \, dx$ be the standard normal cdf evaluated at $x$. If we fix a point $\hat{\theta}$, the expected loss is:

$$
\mathbb{E}[\ell(\hat{\theta}, \theta; c)] = \int \ell(\hat{\theta}, \theta; c) p(\theta) \, d\theta
= \int_{-\infty}^{\hat{\theta}} c\phi(\theta) \, d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})^2 \phi(\theta) \, d\theta.
$$

The first term is proportional to the standard normal cdf $\Phi$ evaluated at $\hat{\theta}$:

$$
\int_{-\infty}^{\hat{\theta}} c\phi(\theta) \, d\theta = c\Phi(\hat{\theta}).
$$

We may also compute the second integral using the following antiderivatives:

$$
\int \theta\phi(\theta) \, d\theta = -\phi(\theta) + C; \quad \int \theta^2\phi(\theta) \, d\theta = \Phi(\theta) - \theta\phi(\theta) + C.
$$

Using the fundamental theorem of calculus, we may use these to calculate

$$
\int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})^2 \phi(\theta) \, d\theta = (\hat{\theta}^2 + 1)(1 - \Phi(\hat{\theta})) - \hat{\theta}\phi(\hat{\theta}).
$$

Finally, the entire expected loss is

$$
\mathbb{E}[\ell(\hat{\theta}, \theta; c)] = (c - \hat{\theta}^2 - 1)\Phi(\hat{\theta}) - \hat{\theta}\phi(\hat{\theta}) + \hat{\theta}^2 + 1.
$$

The derivative with respect to $\hat{\theta}$ is

$$
\frac{\partial}{\partial \hat{\theta}} \mathbb{E}[\ell(\hat{\theta}, \theta; c)] = 2\hat{\theta}(1 - \Phi(\hat{\theta})) + (c - 2)\phi(\hat{\theta}).
$$

Unfortunately, I do not believe we may find a root explicitly, so we would have to rely on numerical root finding. For $c = 1$, the minimal expected loss is achieved at $\hat{\theta} = 0.612$; for $c = 10$, the Bayes action is $\hat{\theta} = -1.0615$; for $c = 100$, the Bayes action is $\hat{\theta} = -2.1167$. In general, the larger $c$, the smaller your estimate should be, due to the potentially high cost of overestimation. For $c = 0$, there is no Bayes action, because we may continue to decrease the expected loss when taking $\hat{\theta} \to \infty$ (there is no reason not to!).

---

4. (Maximum-likelihood estimation.) Suppose you flip a coin with unknown bias $\theta$, $\Pr(x = H) = \theta$, three times and observe the outcome HHH. What is the maximum likelihood estimator for $\theta$? Do you think this is a good estimator? Would you want to use it to make predictions?

Consider a Bayesian analysis of $\theta$ with a beta prior $p(\theta \mid \alpha, \beta) = \beta(\theta \mid \alpha, \beta)$. What is the posterior mean of $\theta$? What is the posterior mode? Consider $(\alpha, \beta) = (1/5, 1/5)$. Plot the posterior density in this case. Is the posterior mean a good summary of the distribution?

**Solution**

The likelihood of HHH is proportional to $\theta^3$, which over the domain $\theta \in [0, 1]$ is maximized at $\theta = 1$. Whether you think this is a good estimator is subjective; however, I certainly wouldn’t use it for prediction, because it completely discounts the (in my opinion, still rather likely) possibility of a tails event.

With a $\beta(\alpha, \beta)$ prior, the posterior is $\beta(\alpha + 3, \beta)$. We may calculate the mean of an arbitrary beta distribution:

$$E[\theta \mid \alpha, \beta] = \int \theta \beta(\theta \mid \alpha, \beta) \, d\theta = \frac{1}{B(\alpha, \beta)} \int_0^1 \theta (\theta^{\alpha-1} (1-\theta)^{\beta-1}) \, d\theta = \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)} = \frac{\alpha}{\alpha + \beta}.$$

Therefore the posterior mean is $\frac{\alpha + 3}{\alpha + 5}$. The posterior mode is $\frac{\alpha + 2}{\alpha + \beta + 1}$. Note that the posterior mode only exists for $\alpha, \beta > 1$.

The prior and posterior distributions are plotted below. An interesting point about the prior is that its mean and median are both $1/2$, but this is simultaneously the anti-mode! It’s an unusual estimator.

![Figure 1: The prior and posterior distributions over $\theta$ for the coin-flipping problem.](http://goo.gl/zigQER -- this is how I would answer this type of question!)
5. (Gaussian with unknown mean.) Let \( x = \{ x_i \}_{i=1}^{N} \) be independent, identically distributed real-valued random variables with distribution \( p(x_i \mid \theta) = N(x_i; \theta, \sigma^2) \). Suppose the variance \( \sigma^2 \) is known but the mean \( \theta \) is unknown with prior distribution \( p(\theta) = N(\theta; 0, 1^2) \).

- What is the likelihood of the full observation vector \( p(x \mid \theta) \)?
- After observing \( x \), what is the posterior distribution of \( \theta \), \( p(\theta \mid x, \sigma^2) \)? (Note: you might find it more convenient in this case to work with the \textit{precision} \( \tau = \sigma^{-2} \).)
- Interpret how the posterior changes as a function of \( N \). What happens if \( N = 0 \)? What happens if \( N \to \infty \)? Does this agree with your intuition?

\textbf{Solution}

In the first part, we use the definition of independence:

\[
p(x \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta) = \prod_{i=1}^{N} N(x_i; \theta, \sigma^2).
\]

We will consider the slightly more general case with an arbitrary Gaussian prior on \( \theta \): \( p(\theta \mid m, s^2) = N(\theta; m, s^2) \). For convenience, we will parameterize the Gaussians on the \( \{ x_i \} \) and \( \theta \) with the \textit{precision} parameters \( \tau = \sigma^{-2} \); \( t = s^{-2} \).

By Bayes’ theorem, we have

\[
p(\theta \mid x) \propto p(x \mid \theta)p(\theta) = \prod_{i=1}^{N} N(x_i; \theta, \tau)N(\theta; m, t) \propto \exp \left( -\frac{1}{2} \left( \tau \sum_{i=1}^{N} (x_i - \theta)^2 + t(\theta - m)^2 \right) \right).
\]

This is an exponentiated, negative quadratic function of \( \theta \) and is therefore proportional to a Gaussian distribution over \( \theta \). Our goal is to identify the mean and precision of this distribution.

We expand:

\[
\tau \sum_{i=1}^{N} (x_i - \theta)^2 + t(\theta - m)^2 = (\tau N + t)\theta^2 - 2(\tau \sum x_i + tm)\theta + (\tau \sum x_i^2 + tm^2).
\]

Notice that \( \sum x_i = N\bar{x} \), where \( \bar{x} \) is the sample mean of the measurements. Notice also that the last term is a constant independent of \( \theta \), which will be absorbed into the normalizing constant. To arrive at a more-familiar form, we “complete the square”: 

\[
(\tau N + t)\theta^2 - 2(\tau N\bar{x} + tm)\theta = (\tau N + t) \left( \theta - \frac{\tau N\bar{x} + tm}{\tau N + t} \right)^2 + c,
\]

where \( c \) is another constant independent of \( \theta \). Therefore:

\[
p(\theta \mid x) \propto \exp \left( -\frac{\tau}{2}(\theta - \theta')^2 \right),
\]

which is a Gaussian distribution with mean and precision

\[
m' = \frac{\tau N\bar{x} + tm}{\tau N + t}; \quad t' = \tau N + t,
\]

\[
\textbf{Solution}
\]
respectively. We recognize that the new precision is the sum of the precisions of each measurement (the $N$ independent measurements $\{x_i\}$ with precision $\tau$ plus one additional prior “measurement” with precision $t$). The posterior mean can be recognized as a precision-weighted average of the measurements, including the “measurement” at the prior mean $m$.

As $N \to \infty$, the sample mean $\bar{x}$ dominates, and the influence of the prior diminishes to zero. If $N = 0$, we rely only on our prior knowledge.
6. (Spike and slab priors.) Suppose \( \theta \) is a real-valued random variable that is expected to either be near zero (with probability \( \pi \)) or to have a wide range of potential values (with probability \( (1 - \pi) \)). Such scenarios happen a lot in practice: for example, \( \theta \) could be the coefficient of a feature in a regression model. We either expect the feature to be useless for predicting the output (and have a value close to zero) or to be useful, in which case we expect a value with larger magnitude but can’t say much else.

A common approach in this case is to use a so-called spike and slab prior. Let \( f \in \{0, 1\} \) be a discrete random variable serving as a flag. We define the following conditional prior:

\[
p(\theta | f, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) = \begin{cases} 
N(\theta; 0, \sigma_{\text{spike}}^2) & f = 0 \\
N(\theta; 0, \sigma_{\text{slab}}^2) & f = 1,
\end{cases}
\]

where \( \sigma_{\text{spike}} \) is the width of a narrow “spike” at zero, and \( \sigma_{\text{slab}} > \sigma_{\text{spike}} \) is the width of a “slab” supporting values with larger magnitude.

In practice, we will never observe the flag variable \( f \); instead, we must infer it or marginalize it, as required.

- Suppose we choose a prior \( \Pr(f = 1) = \pi = \frac{1}{2} \), expressing no a priori preference for the spike or the slab. What is the marginal prior \( p(\theta | \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) \)? Plot the marginal prior distribution for \( (\sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) = (1^2, 10^2) \).

- Suppose that we can make a noisy observation \( x \) of \( \theta \), with distribution \( p(x | \theta, \omega^2) = N(x; \theta, \omega^2) \), with known variance \( \omega^2 \). Given \( x \), what is the posterior distribution of the flag parameter, \( \Pr(f = 1 | x, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2, \omega^2) \)? Plot this distribution as a function of \( x \). What observation would teach us the most about \( f \)? What teaches us the least?

- Given an observation \( x \) as in the last part, what is the posterior distribution of \( \theta \), \( p(\theta | x, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2, \omega^2) \)? (Hint: use the sum rule to eliminate \( f \) and use the result above.)

- Suppose the noise variance is \( \omega^2 = 0.1^2 \) and we make an observation \( x = 3 \). Plot the posterior distribution of \( \theta \), using the parameters from the first part.

**Solution**

For the first part, we use the sum rule:

\[
p(\theta | \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) = \sum_f \Pr(f) p(\theta | f, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) = \frac{1}{2} N(\theta; 0, \sigma_{\text{spike}}^2) + \frac{1}{2} N(\theta; 0, \sigma_{\text{slab}}^2).
\]

The prior density is plotted below.

For the second part, we have from Bayes’ theorem that:

\[
\Pr(f = 1 | x, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2, \omega^2) = \frac{p(x | f = 1, \sigma_{\text{slab}}^2, \omega^2) \Pr(f = 1)}{\sum_f p(x | f, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2, \omega^2) \Pr(f)}.
\]

We must derive the likelihood \( p(x | f, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2) \). We apply the sum rule:

\[
p(x | f, \sigma_{\text{spike}}^2, \sigma_{\text{slab}}^2, \omega^2) = \int p(x | \theta, \omega^2) p(\theta | f, \sigma_{\text{slab}}^2, \omega^2) d\theta.
\]
Figure 2: The prior distribution over $\theta$ for the spike-and-slab prior problem.

For $f = 1$, this becomes:

$$p(x \mid f = 1, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \omega^2) = \int p(x \mid \theta, \omega^2)p(\theta \mid f = 1, \sigma^2_{\text{slab}}, \omega^2) \, d\theta = \int \mathcal{N}(x; \theta, \omega^2)\mathcal{N}(\theta; 0, \sigma^2_{\text{slab}}) \, d\theta = \mathcal{N}(x; 0, \sigma^2_{\text{slab}} + \omega^2).$$

A similar expression holds for $f = 0$. The final result is

$$\Pr(f = 1 \mid x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \omega^2) = \frac{\mathcal{N}(x; 0, \sigma^2_{\text{slab}} + \omega^2)}{\mathcal{N}(x; 0, \sigma^2_{\text{spike}} + \omega^2) + \mathcal{N}(x; 0, \sigma^2_{\text{slab}} + \omega^2)}.$$

This value is plotted as a function of $x$ on the range $x \in [-10, 10]$ below for $\omega^2 = 0.1^2$. Extreme values of $x$ teach us the most, because we can conclude with near certainty that the observation came from the slab. The observations that would teach us the least are $x \approx \pm 2.165$, either of which result in a posterior slab probability of $1/2$ – the same as our prior!

For the third part, we use the sum rule:

$$p(\theta \mid x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \omega^2) = \sum_f p(\theta \mid f, x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \omega^2) \Pr(f \mid x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \omega^2).$$

We need to compute the posterior on $\theta$ given $f$ and $x$. Here we may use the result from the last problem:

$$p(\theta \mid f = 0, x, \sigma^2_{\text{spike}}, \omega^2) = \mathcal{N}(\theta; \tau^{-1}_{\text{spike}}\omega^{-2}x, \tau^{-1}_{\text{spike}})$$
$$p(\theta \mid f = 1, x, \sigma^2_{\text{slab}}, \omega^2) = \mathcal{N}(\theta; \tau^{-1}_{\text{slab}}\omega^{-2}x, \tau^{-1}_{\text{slab}}).$$
Figure 3: The posterior probability for the slab flag ($f = 1$) for the spike-and-slab prior problem as a function of the observation $x$.

where $\tau_{\text{spike}} = (\omega^{-2} + \sigma_{\text{spike}}^{-2})$, and $\tau_{\text{slab}}$ is defined similarly.

Finally, the posterior for $\theta$ for the observation $x = 3$ with $\omega^2 = 0.1^2$ is plotted below.
Figure 4: The posterior density for the parameter $\theta$ given the example observation $x = 3$, $\omega^2 = 0.1^2$. 